Elements of Semantic Web Infrastructure for Maritime Information

Raphael Malyankar
Department of Computer Science and Engineering
Arizona State University

BIOGRAPHY

After earning his doctorate in computer science, the author has been employed as a research scientist in the Department of Computer Science and Engineering at Arizona State University. He has also worked as software engineer in manufacturing scheduling and credit card billing. He earned a Master's in computer science from the University of New Hampshire and a doctorate in computer science from Arizona State University. His research interests include web technologies including semantic web research, resource allocation and planning, knowledge representation, and intelligent agents.

ABSTRACT

This paper describes elements of Semantic Web infrastructure for maritime information, including navigation information. The work is motivated by the need to prepare maritime information representation and distribution for future-generation Web technology, especially the Semantic Web. The infrastructure elements described here consist of computational ontologies and a prototype XML-based markup language for maritime information. Computational ontologies and markup languages for a domain are both essential for developing Semantic Web applications for that domain.

The first phase of the work consists of the construction of a computational ontology for navigation information and nautical chart symbology. (A computational ontology for a domain consists of a collection of concepts defined in the domain and the relationships between these concepts, expressed in a form that can be processed by software.) Ontological information is acquired from multiple sources, including standards documents, database schemas, lexicons, collections of symbology definitions. The sources of ontological knowledge and the contribution of each source to the overall ontology are described in this paper. In the second phase, the computational ontology is used to create a prototype of the language - Maritime Information Markup Language (MIML) - for tagging documents within the maritime domain. An overview of this prototype markup language is included.

The use of this markup language and ontology in a demonstration application is then described. One application area for these information infrastructure elements described here is the integrated retrieval of maritime information from diverse sources, ranging from Web sites to nautical chart databases and text documents. An architecture for such a retrieval system is described. A web-based demonstration prototype based on these concepts has been developed.

The work performed to date demonstrates the utility of the computational ontology as a multi-level index to diverse information sources, and in creating a uniform interface for information retrieval. The use of the MIML prototype in markup of, and extraction from, text documents is also demonstrated. The computational ontology and markup language are enabling technlogy for future smarter information retrieval and reasoning systems, for example, automated passage planning and intelligent navigation software.

INTRODUCTION

The Semantic Web is a paradigm for Web content and processing that attempts to introduce 'meaning' to the World Wide Web, in the sense of making it easier for Web applications to process and react to the *content* of documents, and not just the presentation and markup. Berners-Lee, Hendler, and Lassila describe it as "not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [22]. A visionary application involving the *automated* discovery of information about medical providers and making a plan for a visit to a selected provider is described in that paper.

The implications of the Semantic Web in the maritime domain lie in enabling smarter processing of the voluminous information — from different sources and in different forms (text, diagrams, etc.) — that may need to be dealt with in the course of a trip. This information processing involves, in addition to course plotting and information about navigation aids and hazards, weather reports, data about the nature and status of facilities in relation to the nature and purpose of the voyage (e.g., the information required for a supertanker is different from that required for a sailboat), etc.

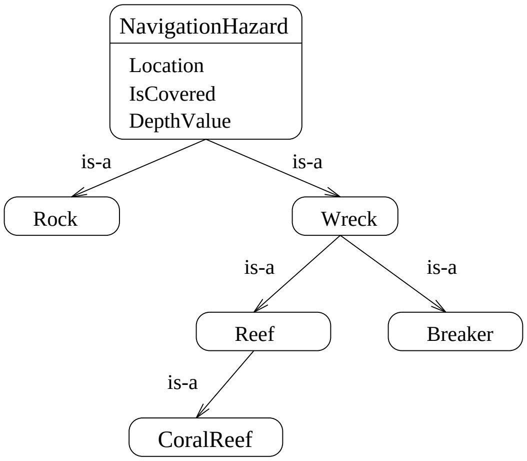


Figure 1: Part of the taxonomy for hazards

Berners-Lee, Hendler, and Lassila [22] mention knowledge representation, ontologies and agents as core technology for the semantic web. In artificial intelligence, an *ontology* is a set of definitions of the concepts that exist in a particular domain and the relationships between them. A *computational ontology* is a a collection of terms, formal definitions, and constraints, which can be processed by software, and which increases the scope of computational methods applied to the relevant domain. An ontology for a domain can be considered as an extended 'information model' for the domain. A basic component of any ontology is a taxonomy or classification of concepts — an *is-a* hierarchy or class hierarchy. Figure 1 shows part of the hierarchy for navigation hazards.

This paper describes the creation of an ontology (or information model) for maritime information, and the use of this information model in markup and integrated retrieval from multiple sources. The creation of a model of available knowledge and information constitutes 'ontological engineering'. For our project, ontological information is acquired from multiple sources, including standards documents, database schemas, lexicons, collections of symbology definitions, and also from semi-structured documents.

Another essential component is one or more markup languages for describing document content. The World-Wide Web Consortium's Extensible Markup Language (XML) [23] is currently the most relevant and mature framework for such a markup language. The computational ontology created is being used to create an prototype XML-based markup language (Maritime Information Markup Language — MIML) for tagging documents within this domain. This markup language is used (as an initial demonstration of the kind of application that will be enabled) in a Web-based information retrieval system that obtains information from Web sites, digital nautical charts, and marked-up text documents.

The next section describes the sources of ontological knowledge and the contribution of each source to the overall ontology. An overview of a markup language for maritime information derived from this ontology follows. The use of this markup language and ontology in the demonstration application is then described.

ONTOLOGY CONSTRUCTION

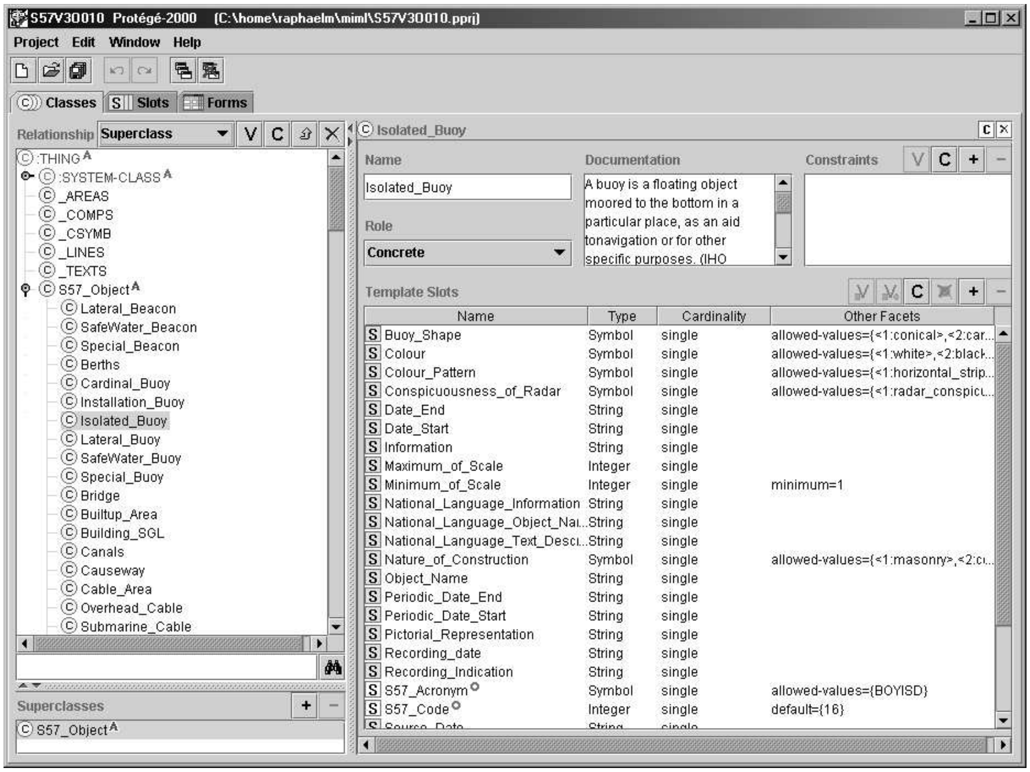
In gathering ontological knowledge, a deliberate decision was made to use standardized sources wherever possible, with the dual purpose of leveraging the prior work of domain experts and enhancing acceptability of the resulting product. The following sources were used:

Standards Documents: A normative standard for digital nautical chart content is the IHO (International Hydrographic Organization) S-57 Transfer Standard for Digital Hydrographic Data [11]. The 'object catalog' section of this document consists of a list of chart entities, definitions, and entity attributes, which gives us a collection of domain entities that can be considered canonical as far as the scope of the standard goes. Extraction from this 'object catalog' was automated using graph traversal algorithms that exploit links between entities and attributes. The automated extraction resulted in 173 classes. A comparison of 10% (selected at random) of the extracted information with the original source indicated error rates of 8% to 20% (for different categories of ontological knowledge — classes/types/attributes). The additional effort needed to reduce this rate in the automated extraction was not undertaken, as it proved no very laborious task to make the corrections by hand (about 10 hours for a non-expert who compared the extracted ontology with the original source). Figure 2 shows part of the ontology derived from the IHO S-57 standard.

The Spatial Data Transfer Standard [6] was another source. The parts we used were the list of 'included terms' (analogous to a hyponym list) and attribute definitions. Extraction from this was less satisfactory in some ways, since these sections are less rigorous than the object catalog of the S-57 standard, but, on the other hand, the lists cover more of the terms used in practice.

Databases: The primary digital chart database we have used so far is the set of sample Digital Nautical Chart (DNC) data files available from the National Imagery and Mapping Agency (NIMA), covering the San Diego Harbor and approaches. The DNC database has somewhat more *semantic* structure than the aforementioned standards, consisting as it does of feature classifications organized by 'layers', for example, environmental features, cultural features, land cover features, etc. Induction of ontological knowledge from this consisted of mapping the structure to a class hierarchy. Taxonomical information that could be directly extracted from the table names in this database therefore consists of relationships between the abovementioned features/classes. Approximately 134 classes were mined from this database.

As with the S-57 standard, this database and schema covers only chart entities, and the terminology is even more restricted (and to some extent, more linguistically opaque) than the S-57 standard, due to the use of abbreviated names for entities and attributes, and the lack of textual definitions.



:THING
:SYSTEM-CLASS
- _AREAS
- _COMPS
- _CSYMB
- _LINES
- _TEXTS

S57_Object
- Lateral_Beacon
- SafeWater_Beacon
- Special_Beacon
- Berths
- Cardinal_Buoy
- Installation_Buoy
- Isolated_Buoy
- Lateral_Buoy
- SafeWater_Buoy
- Special_Buoy
- Bridge
- Builtup_Area
- Building_SGL
- Canals
- Causeway
- Cable_Area
- Overhead_Cable
- Submarine_Cable

Superclasses: S57_Object

Isolated_Buoy

Name: Isolated_Buoy

Role: Concrete

Documentation: A buoy is a floating object moored to the bottom in a particular place, as an aid tonavigation or for other specific purposes. (IHO

Name	Type	Cardinality	Other Facets
Buoy_Shape	Symbol	single	allowed-values={<1:conical>,<2:car...
Colour	Symbol	single	allowed-values={<1:white>,<2:black...
Colour_Pattern	Symbol	single	allowed-values={<1:horizontal_strip...
Conspicuousness_of_Radar	Symbol	single	allowed-values={<1:radar_conspicu...
Date_End	String	single	
Date_Start	String	single	
Information	String	single	
Maximum_of_Scale	Integer	single	
Minimum_of_Scale	Integer	single	minimum=1
National_Language_Information	String	single	
National_Language_Object_Na...	String	single	
National_Language_Text_Desc...	String	single	
Nature_of_Construction	Symbol	single	allowed-values={<1:masonry>,<2:c...
Object_Name	String	single	
Periodic_Date_End	String	single	
Periodic_Date_Start	String	single	
Pictorial_Representation	String	single	
Recording_date	String	single	
Recording_Indication	String	single	
S57_Acronym	Symbol	single	allowed-values={BOYISD}
S57_Code	Integer	single	default={16}
Source_Date	String	single	

Figure 2: Ontology from S-57 Hydrographic Information Standard in *Protege*

Lexicons and Symbology Definitions: We also used the Stanford Medical Informatics group *Protégé* tool [7] with the standard collection of chart symbology published by the National Oceanic and Atmospheric Administration (NOAA *Chart No. 1*) [17] to create an ontology of navigation aids, hazards, and other entities. *Chart No. 1* provides brief descriptions of the meaning of each symbol. It is organized semantically (in that related symbols are in the same section or subsection). Approximately 500 classes were created from this source. Definitions available within this document were supplemented by using *Chapman Piloting* [13] and an online dictionary of chart terms (discovered and used by the knowledge entry personnel, who were computer science students unfamiliar with nautical terms). Ontology creation based on these documents consisted of manual entry of information using Protégé, due to the lack of electronic versions of the symbology definitions.

Semi-structured material: The *United States Coast Pilot* consists of 'lightly structured' text, with each volume containing a preliminary chapter containing navigation regulations (including a compendium of rules and regulations, specifications of environmentally protected zones, restricted areas, etc.), followed by chapters dealing with successive sectors of the coast. Each chapter is further divided into sections (still in geographical order); each section is further divided into sub-sections and paragraphs describing special hazards, recognizable landmarks, facilities, etc.

The internal structure of subsections and paragraphs provides taxonomic hints, indicating, for example, which leaf entities are categorizable as sub-classes of weather conditions, as well as providing a small amount of additional taxonomical information that extends taxonomies derived from other classes. The *Coast Pilot* was selected as being considered normative (in the sense of using well-understood terms) and comprehensive. A version marked up with XML would have proved invaluable for ontology learning, but there is no such version available at this time (indeed, it is the aim of this project to create such a marked up version).

There is a certain amount of overlap in the ontological information derived from the different sources above, in addition to structure mismatches (for example, information that is denoted as an attribute in one ontology may be used for sub-classing in another). The reuse of other computational ontologies was explored early on, but they (e.g., the SENSUS ontology in Wordnet turned out to define terms in usages that are either not relevant, or 'wrong' in the maritime information domain — for example, a 'bridge' is a trafficable passageway for land travel, but an obstruction or landmark for waterborne vessels.

We have discovered that though there is a certain amount of duplication between the above sources, they are largely independent and produce different parts of the taxonomy for the maritime information domain as a whole. It is therefore necessary to merge the 'sub-ontologies', i.e., construct a larger ontology that contains all the terms and that reconciles discrepancies between them. An initial merger was done using a Perl program; a Protégé plugin for this task is being developed. The theoretical underpinning of our approach to the merging problem is described elsewhere [14]. Ontology merging and ontology mapping (relating concepts in one ontology to those in another) are active areas of research [3; 9; 18; 19].

When the ontological knowledge has been compiled, the bulk of markup language definition work can begin. This is described next.

MARKUP LANGUAGE

This section briefly describes a seed, skeletal version of the Maritime Information Markup Language (MIML), with the intent of illustrating how the markup language could be constructed from basic concepts in knowledge representation (i.e., ontologies). The version described here has not been through an external review process, which would be essential for a deployable version, and should not be regarded as authoritative or complete, consisting as it does primarily of views of the author of the current paper.

Foundation of the Language

The markup language is founded on transition from computational ontologies to data types, entities, and attributes. This turns out to be a two-way process - a small part of the sum of ontological knowledge was formalized through doing a sample markup of part of the Coast Pilot *ab initio* (without predetermined elements, attributes, etc.). The bulk of the ontology design work to date, however, was based on other sources (i.e., was generated prior to any markup).

Transition from Ontologies to Markup

The transition from ontology to markup consisted of the following:

- Certain markup tags (elements) were defined by naming them after classes in the ontology. This decision was based on the belief that if a concept was considered important enough and distinct enough to be reified in the ontology (by being named as a distinct class), it is likely to be used as a markup element. The classes so treated were the leaf classes in the ontology (those that are not further subclassed).
- Slots in the ontology were generally mapped to element attributes in the markup language, since in our domain they generally consist of small “atoms” of information such as the name of a place or one of an enumerated set of values (for example, the “nature of sea floor” slot/attribute can have the value “mud”, “clay”, “sand”, “rock”, etc.). The slots were also used to define datatypes in the XML schema, and restrictions on the allowed values for an attribute (where they exist) were maintained by defining the corresponding restrictions on the target type in the schema (e.g., an enumerated type of the base type “string” was defined corresponding to an ontology slot in Protégé of type “symbol” with a facet restricting the possible values of this “symbol”).
- A small number of selected slot values were mapped to elements in the markup language. This applies to those that were felt to be important enough to merit their own notes in the target documentation.

Unsurprisingly, very little relationship between the *is-a* hierarchy in the ontologies (class/subclass relationships) and the element containment relationships in the markup language exists. Part-whole relationships in the ontology do appear to be maintained across the mapping from ontology to markup but this preservation does not always exist, and is naturally uni-directional (i.e., containment in the markup language does not imply a part-whole relationship in the source ontology). This implies that information is being lost in the transition from ontology to markup and vice versa, meaning that both ontologies and markup will be needed for a proper understanding and processing of target documents.

Language Sub-divisions

The markup language appears to divide naturally or intuitively into partitions or different sub-languages, each corresponding to a separate XML schema [25; 26; 27]. The partition is based upon taking into account the source of the ontological knowledge, the use of markup, interdependencies, and expectations for change control. The existence of a specific source for a part of ontological knowledge usually indicates its use within sub-domains of the overall domain - for example, a weather ontology (or markup) will be used by forecasters, distributors of weather information, and consumers (mariners), but the digital charts community is not interested in weather insofar as the making of digital chart databases is concerned. Change control and updating of ontologies and markup will be simplified by limiting these responsibilities to the interested sub-communities.

In practice, this partitioning is expected to be implemented using different namespaces (or another suitable partition mechanism). Figure 3 contains a conceptual overview of the author’s view of a possible partitioning of MIML into different schemas.

The diagram on the left of Figure 3 shows the conceptual structure of MIML itself, with markup (sub)languages at the core consisting of:

1. The S-57 core, comprising entities and attributes described in the International Hydrographic Organization’s S-57 standard and only those entities and attributes.
2. A geography markup component, tentatively identified as GML (Geography Markup Language) which is being prepared by the OpenGIS consortium. This is included primarily to represent low-level primitives such as shapes (lines, polygons, etc.).
3. A communications (sub)language (”Comm” in the figure) for describing communications-related information (VHF channel information, radio call signs, telephone information, etc.).
4. A Services (sub)language, to describe port facilities, small craft repair information, etc.

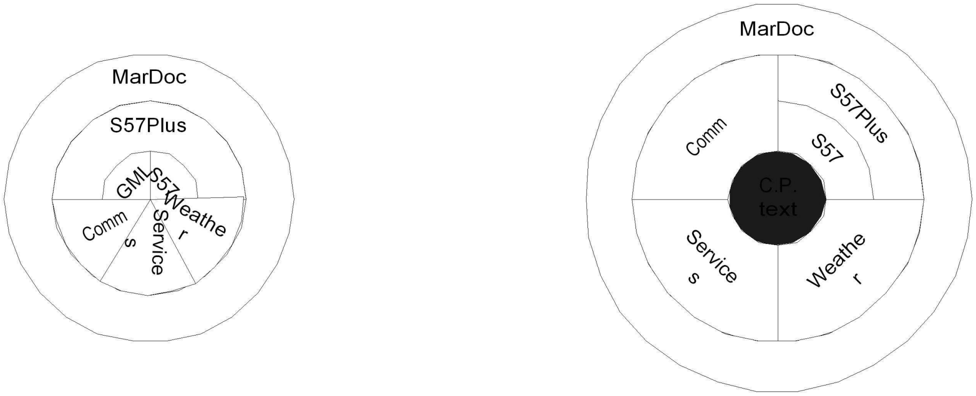


Figure 3: Schemas comprising MIML

5. A Weather (sub)language, to describe wind and sea conditions, weather forecasts, etc.

The outer components in the figure are: a sub-language tentatively called S57Plus, intended to extend the S57 core with markup information that is generally required when S57 elements are discussed in texts, but which is not contained in the S57 standard; and a *MarDoc* component, intended for the markup of document structural elements that are not part of the domain itself, but which recur in target documents (for example, a ”chart” element that could demarcate the part of a text document that contains information pertaining to a specific nautical chart).

Layering in Figure 3 denotes ”use relationships” - for example, elements defined in the ”MarDoc” part are expected to use (contain) elements in all the other parts, but the S57 component is not expected to use elements or definitions from other parts.

The diagram on the right of Figure 3 illustrates the use of markup components with a specific target document or information resource (the shaded part at the center). Note that not all the components need be used to mark up any specific information resource.

The formal definition of the language is envisaged to be in terms of DTDs (Document Type Definitions) and XML schemas; the components of such a definition are described next.

Schemas and DTDs

A DTD has been prepared for a sample ”shoreline” chapter of the *Coast Pilot* (i.e., a chapter other than the preliminary chapter concerning navigation regulations, etc.). Part of this is shown in Figure 4. A DTD for the *Local Notice to Mariners* has also been prepared (omitted in the interests of brevity). XML schemas derived from the S-57 ontology and the Coast Pilot DTD are under development; sample elements for *Wharves* and *Pier* are shown in Figure 5. For most elements in the Coast Pilot schema, the type is ”mixed”, because the document currently consists of a sequence of paragraphs in English; each paragraph either expands on something in the corresponding nautical chart (e.g., adds a note about a specific navigation hazard), or contains condensed information from another document (e.g., about port facilities), or provides location-specific information (e.g., contact information for local authorities, reminders about local regulations, etc.). In other words, the *Coast Pilot* is basically a *text document* that repeats certain forms of expression in a restricted vocabulary of a natural language (English) for different parts of the coastline; it is not, at this time, intended for automated processing. We hope to move towards this goal via the DTDs and schemas described here, but meanwhile it is important to retain its human readability.

Figure 6 shows a marked-up fragment of Volume 7 of the *Coast Pilot*. This fragment is part of Chapter 4, which covers the California coast from San Diego to Point Arguello, and describes, amongst other things, weather, navigation hazards, aids for navigation, local regulations, contact information, harbor facilities, etc. The section in the figure comes from the portion describing harbor facilities in San Diego Bay. (Ellipses denote material left out of this figure for brevity’s sake.)

Given the above infrastructural elements – ontologies and a skeleton markup language – it is possible to construct a proof-of-concept application demonstrating a standardized interface to information retrieval. This is discussed next.

DEMONSTRATION APPLICATION

A ‘passage plan’ is, for the purposes of this project, an answer to the questions: ”How do I get from X to Y? What will I encounter on the way, and what will I find when I get there? What do I need to know for this particular journey”? Passage planning involves not just plotting a safe route, but also includes generating a report about hazards that may be encountered, facilities available along the route and at the destination, weather and tide conditions that may encountered during the voyage, etc. The passage plan depends on the type of vessel and the purpose of the journey, since information that may be of interest to a freighter may be irrelevant to a small pleasure craft. The use of these concepts, and of MIML is demonstrated in a prototype Web-based application.

```
<!DOCTYPE CoastPilot [
 <!ELEMENT CoastPilot (Scope, GeneralDescription, GeneralMaterialOnChart+)>
;
 <!ATTLIST CoastPilot Volume CDATA "" Chapter CDATA "" >
 <!ELEMENT Scope (From, To   )>
 <!ELEMENT From EMPTY>
 <!ELEMENT To EMPTY>
 <!ATTLIST From Name CDATA "" Latitude CDATA "" Longitude CDATA "">
 ...
 <!ELEMENT Pier (#PCDATA | Berth |Dimensions | Service )*>
 <!ELEMENT Berth (#PCDATA | BerthNumber | Dimensions |
                      BerthFacilities | Service  )*>
 <!ELEMENT BerthNumber EMPTY>
 <!ATTLIST BerthNumber No CDATA "">
 <!ELEMENT Dimensions (#PCDATA)>
 ...
 <!ELEMENT Wharves (#PCDATA | Xlink | PierArea )*>
 <!ELEMENT PierArea (#PCDATA | Pier )*>
 ...
```

Figure 4: Partial DTD for the *Coast Pilot*

```
 <element name=’Pier’>
  <complexType mixed=’true’>
   <choice minOccurs=’0’ maxOccurs=’unbounded’>
    <element ref=’Berth’/>
    <element ref=’Dimensions’/>
    <element ref=’Service’/>
   </choice>
   <attribute name=’Name’ type=’string’ use=’default’ value=’’/>
  </complexType>
 </element>
 ...
  <element name=’Wharves’>
  <complexType mixed=’true’>
   <choice minOccurs=’0’ maxOccurs=’unbounded’>
    ...
    <element ref=’PierArea’/>
   </choice>
  </complexType>
 </element>
```

Figure 5: XML Schema elements for the *Wharves* and *Pier* elements

```
<Chart>
 <ChartNumber>18773</ChartNumber> <ChartNumber>18772</ChartNumber>
 <Location> San Diego Bay is 10 miles NW of the Mexican boundary</Location>
 <Description>San Diego Bay is where California’s maritime history...</Description>
 ...
 <Anchorages>General anchorages, special anchorages, and ... </Anchorages>
 <Tides> The mean range of tide is 4.0 feet at San Diego ... </Tides>
 <Currents> The currents set generally in the direction of...</Currents>
 ...
<Wharves>
 The San Diego Unified Port District owns the deepwater commercial facilities in...
 <PierArea>
  <Pier  name ="B Street Pier, Cruise Ship Terminal">
  (32 deg. 43’02"N., 117 deg. 10’28"W.): 400-foot face, 37 to 35 feet alongside;
  1,000-foot N and S sides, 37 to 35 feet alongside;...
  </Pier>
  <Pier  name ="Broadway Pier, S of B Street Pier">
  135-foot face, 35 feet alongside; 1,000-foot N and S sides, 35 feet...
  </Pier> ...
  <Pier  name ="Tenth Avenue Marine Terminal">
   <Berth  name="Berths 1 and 2">
    Concrete bulkhead, 1,170 feet of berthing space; 27 feet alongside...
   </Berth>
   <Berth  name="Berths 3 and 6"> ... </Berth>
   <Berth  name="Berths 7 and 8"> ... </Berth>
  </Pier>
 <PierArea>
</Wharves>
...
</Chart>
```

Figure 6: Marked-up fragment of Chapter 4, Volume 7 of the *Coast Pilot*

Content sources for prototype site

The ‘content’ sources for passage planning are Web sites with real time information, the Coast Pilot, and programs that generate information as and when required. They fall into the following categories:

Static text documents: The primary text document currently being used is the Coast Pilot, described earlier. It includes descriptions of particular items of interest, some of which are also shown in nautical charts (such as lighthouses and beacons), and other descriptions which are either not available in the nautical charts and other places, or not apparent from them, such as special local tidal dangers. (Where information in the CP duplicates that in other sources, we interpret it as emphasizing important features and dangers.) It also contains a few diagrams and photographs taken from a mariner’s point of view, information on anchorages, etc., and pointers to other sources, for example, to the U.S. ‘Port Series’ for more information on facilities at a specific port. Markup of the Coast Pilot with MIML tags is currently being done manually; we hope to automate part of this markup process in the future.

Web sites with real-time information: Certain information is being made available in near-real-time by both official and unofficial sources, especially weather conditions, forecasts, and warnings. The National Databuoy Center (NDBC) Web site provides recent weather data from databuoys all along the U.S. coastline. Certain marinas have also begun putting local conditions on their Web sites. Tide predictions are available from another NOAA site. The prototype incorporates information from databuoys (via the NDBC Web site).

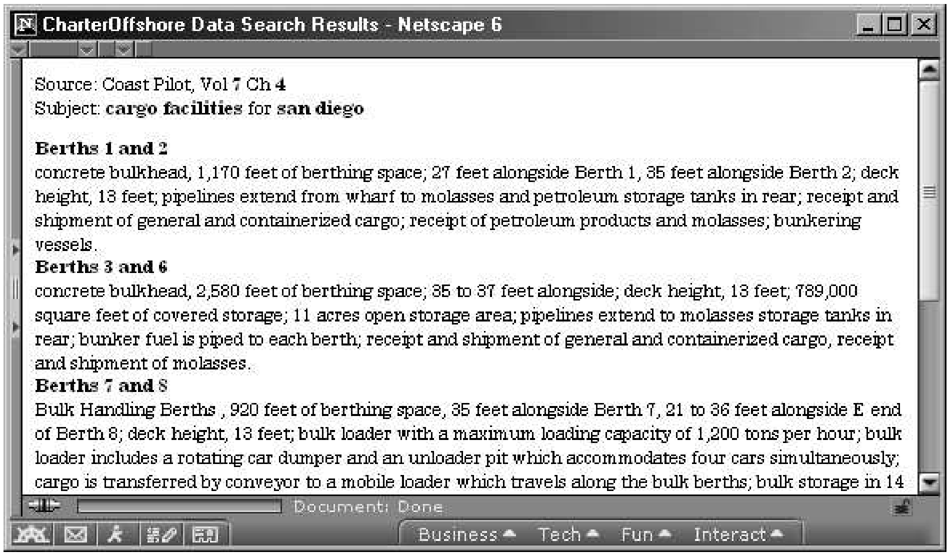
Chart databases: Data extracted from DNCs was loaded into a local database server (due to the difficulty of querying DNC files). The contents are essentially tables of features and their attributes. These tables are easily transformed into an object-relational database form. This source provides information about such items as coastlines, marker buoys, lighthouses, depth measurements, and other nautical chart features.

Dynamically generated content: Certain content (tide predictions) is generated by programs residing on the local web server.

CAPABILITIES DEMONSTRATED BY PROTOTYPE

The capabilities demonstrated in the prototype application are:

Extraction from XML documents: Relevant elements from the marked-up Coast Pilot are extracted in response to



Source: Coast Pilot, Vol 7 Ch 4
Subject: **cargo facilities** for **san diego**

Berths 1 and 2
concrete bulkhead, 1,170 feet of berthing space; 27 feet alongside Berth 1, 35 feet alongside Berth 2; deck height, 13 feet; pipelines extend from wharf to molasses and petroleum storage tanks in rear; receipt and shipment of general and containerized cargo; receipt of petroleum products and molasses; bunkering vessels.
Berths 3 and 6
concrete bulkhead, 2,580 feet of berthing space; 35 to 37 feet alongside; deck height, 13 feet; 789,000 square feet of covered storage; 11 acres open storage area; pipelines extend to molasses storage tanks in rear; bunker fuel is piped to each berth; receipt and shipment of general and containerized cargo, receipt and shipment of molasses.
Berths 7 and 8
Bulk Handling Berths , 920 feet of berthing space, 35 feet alongside Berth 7, 21 to 36 feet alongside E end of Berth 8; deck height, 13 feet; bulk loader with a maximum loading capacity of 1,200 tons per hour; bulk loader includes a rotating car dumper and an unloader pit which accommodates four cars simultaneously; cargo is transferred by conveyor to a mobile loader which travels along the bulk berths; bulk storage in 14

Figure 7: Elements describing cargo facilities, extracted from the *Coast Pilot*

a user query. Relevance is judged based on proximity to the location(s) specified (and the route between source and destination), the type of vessel and purpose of the voyage, and in response to the optional question mentioned earlier. Figure 7 shows the response to a question about cargo facilities at San Diego.

Real-time information presentation: Information about weather conditions as reported by NDBC data buoys is retrieved from the NDBC web site, and processed into a form suitable for presentation, this time with MIML markup added automatically.

Data retrieval: The databases created from DNCs are queried with SQL queries and the results transformed into forms suitable for presentation. This transformation currently involves statistical post-processing of the information, the nature of this post-processing depends on the form of the query, especially when the retrieval is raw material for an response to the optional user questions mentioned earlier. Transformation is discussed further in the next paragraph.

Single interface for simple question answering: The prototype is able to answer questions posed using a limited vocabulary and syntax. Questions can be asked in ways that are close to natural language (e.g., "can I anchor . . ."). Pattern-matching is used to transform this natural-language question into a query that can be executed by the database back-end. The techniques used in information retrieval and processing of retrieved information may involve the capabilities described earlier in this section. In some cases, the raw data retrieved from the database undergoes post-processing depending on the form of the question; for example, the questions "show the sea floor off . . ." and "can I anchor near . . ." both retrieve the same raw data (sea floor characteristic data points), but the first form produces a table of values, while the second combines the retrieved values into an assessment of the general sea floor description in the same location.

The primary interface with the user consists of a form to be filled out with information about the journey, including the location (either source-destination or a single point), type of vessel (cargo, sail, etc.), time of journey, and, optionally, specific questions about such items as anchorages, local facilities, depths, etc. The Web server transforms the form into a collection of sub-queries, each formulated for the specific knowledge sources available (here, Web sites, DNC database, marked-up CP files, and a tide prediction program).

The current version of this page limits its search to the four sources mentioned earlier and answers a limited range of questions, being constrained by the limited richness of structure of the sources (e.g., the CP is marked up with MIML tags only at sentence- or paragraph-level detail, whereas tagging parts of sentences is required for more sophisticated retrieval). Efforts to integrate more sources and enhance the expressiveness of the markup language are underway.

The schematic in Figure 8 shows the conceptual structure of the demonstration application. Figure 8 shows a question entered by the user in (semi) natural language being transformed to use standard terms by a black-box language processor. A multi-level index derived from computational ontologies for the domain is used to re-cast the question into knowledge-source specific formats which are dispatched to the appropriate information sources (the filled circles). The information sources may be wrapped in one or more markup languages (the unshaded arcs) or unshielded (not use any markup at all) as in the leftmost circle in the figure.

The demonstration application in its current stage of development does not deal with the route-planning problem ("how do I get from X to Y"), because similar issues have long been addressed in path planning research within artificial intelligence, and the computational magnitude of this

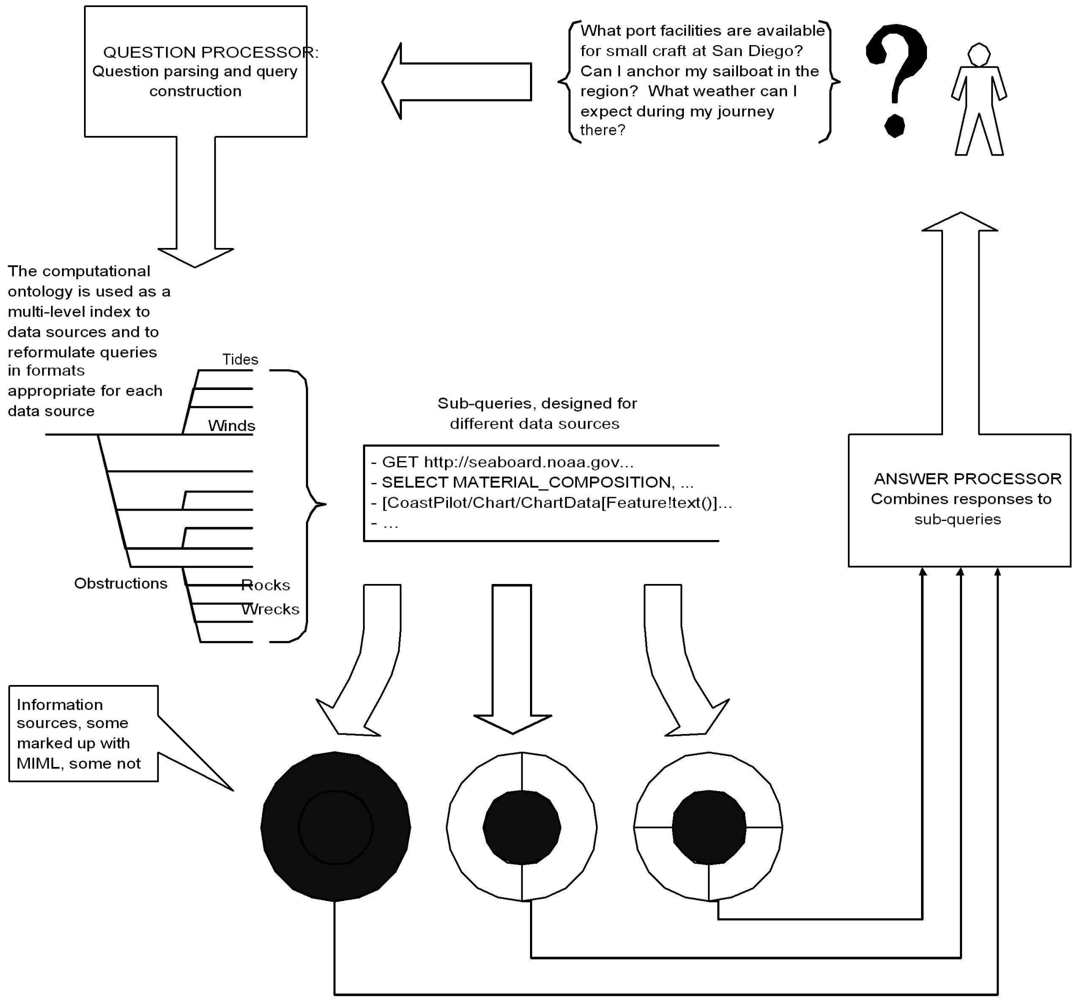

QUESTION PROCESSOR:
Question parsing and query construction
What port facilities are available for small craft at San Diego? Can I anchor my sailboat in the region? What weather can I expect during my journey there?
?
The computational ontology is used as a multi-level index to data sources and to reformulate queries in formats appropriate for each data source
Tides
Winds
Obstructions
Rocks
Wrecks
Sub-queries, designed for different data sources
- GET http://seaboard.noaa.gov...
- SELECT MATERIAL_COMPOSITION, ...
- [CoastPilot/Chart/ChartData[Feature!text()]...
- ...
ANSWER PROCESSOR
Combines responses to sub-queries
Information sources, some marked up with MIML, some not

Figure 8: Schematic for Information Retrieval Application

particular problem prevented anything more than a superficial solution with available resources. It does attempt to deal with the other components of what we call the 'passage planning question'. Note that the passage plan depends on the context of the question, especially type of vessel and the purpose of the journey, since information that may be of interest to a freighter may be irrelevant to a small pleasure craft.

RELATED WORK

A significant amount of data for the waterborne transport community is already being distributed via digitized methods, including the World-Wide Web. Specific application areas in the maritime domain for the concepts described here domain are described by Spalding and Pirzada [20] and by Spalding and others [21]. Certain next-generation information systems that are being planned, for example, the Marine Data Hub, augmented-reality navigation aids, vessel traffic management systems, and Intelligent Waterways System mentioned in these papers, will need efficient processing and integration of multiple kinds of information, from multiple sources, not all of which currently use the same data model. Linking of ontologies for these disparate sources and kinds of information, using commonly known schemas and markup languages should enhance the capabilities of these systems. A foundation for future intelligent navigation aids and software will be provided by infrastructure elements of the kind described in this paper.

On the theoretical side, concerning information gathering, Knoblock and Ambite [12] describe the use of a domain model in formulating queries for different knowledge sources represented by different agents. Noy and Musen [18; 19] describe an algorithm and tool for merging ontologies in Protégé. Chalupsky [2] describes OntoMorph, a tool for translating symbolic knowledge from one KR formalism to another, and describes ontology alignment in [3]. Hovy [9] describes a procedure for ontology alignment and heuristics for suggestions, including pattern matching on strings, hierarchy matching and data/form heuristics. Ontology analysis and merging in *Chimæra* is described in [15]; the techniques used include syntactic analysis of class and slot names, taxonomic resolution, and semantic evaluation (for example, slot/value type checking and domain-range mismatches).

The work in deriving markup languages from ontologies that is most closely related to that described here appears to be that of Erdmann and Studer [4], which describes derivation of DTDs from ontologies by mapping concepts and attributes to XML elements. A tool that does this mapping ("DTDmaker") is described. In purpose and function, it appears to be similar to the techniques used by us to produce XML schemas from our ontologies, except that we produce schemas instead of DTDs. Erdmann and Studer claim that both DTDs and ontologies are needed for document understanding and processing, reinforcing the observation in this paper that there is information in each that is not reproduced in the other. Hunter [10] describes a similar standard-based approach to ontology and markup definition, for MPEG-7. Mitra, Wiederhold, and Decker [16] describe the integration of heterogeneous information sources based on conversion of disparate conceptual models to a single conceptua l model, similar in intent to what is being done in the prototype retrieval application described here, but apparently different in the emphasis on the reconciliation of conceptual models for interoperability. Euzenat [5] describes a formal approach to similar problems. RDF and RDFS [24] are even more closely related to ontologies and knowledge representation than XML. DAML (DARPA Agent Markup Language) [8; 1] is intended as an extension to XML and RDF, and is another alternative for semantic descriptions. The work on RDF and RDFS has given rise to some controversy in the Web and knowledge representation communities, and DAML is a recent development, intended to provide a system that is more rigorously specified and is more amenable to processing by software based on the principles of description logic. Future work is expected to move to these or other alternatives that are semantically richer than XML. However, a significant amount of research and development remains to be done before either RDF/RDFS or DAML can be used in a deployable application.

SUMMARY AND FUTURE WORK

The primary purpose of this paper was describing the creation of a markup language for maritime information (MIML) from computational ontologies and the use of this language in information retrieval from documents and other sources used in the field. MIML is still in the early stages of development, and needs to go through a standards process before it can gain wide acceptance in the field. Explorations of the applicability of this research to the proposed Waterway Information Network [20; 21] are planned. Research plans for the future include demonstrating capabilities beyond information retrieval, especially intelligent reasoning, using the retrieval and access capabilities provided by markup; this will involve 'drill-down' markup, to lower levels than in the sample fragment of Figure 6. Querying of large databases of XML documents, the use of markup and ontologies in delivery of information to users, and the use of markup in updating databases (and documents) and in translating between heterogeneous databases will be investigated. The implications for our research of the Resource Description Framework (RDF) and DARPA Agent Markup Language (DAML), which are both currently still under development, will also be examined. Incorporation of a high-level conceptual model and description logic-based reasoning will also be explored.

ACKNOWLEDGEMENTS

The efforts of Koi-Sang Leong, Helen Wu, and Rashmi K Iyengar in creating the ontology and markup of the Coast Pilot are gratefully acknowledged, as are the efforts of Suresh Lalwani and Jayagowri Renuka in programming for the demonstration described. The help of personnel from the USCG, especially Kathy Shea and Jay Spalding, and of Oren Stembel of NOAA is also gratefully acknowledged. This work was partially supported by the Digital Government Program of the National Science Foundation under grant EIA-9983267, NOAA, Maptech, and the U.S. Coast Guard. The opinions, findings, conclusions and recommendations expressed in this material are those of the author and do not necessarily reflect the views of these entities.

NOTES

The demonstration information retrieval application mentioned in this paper is available on the WWW via a link from the project Web page (www.eas.asu.edu/∼gcss/research/navigation/).

References

[1] Mark Burstein, Jerry Hobbs, Ora Lassila, David Martin, Sheila McIlraith, Srini Narayanan, Massimo Paolucci, Terry Payne, Katia Sycara, and Honglei Zeng. DAML-S: Semantic markup for web services, May 2001. www.daml.org/services/daml-s/2001/05/daml-s.html.

[2] H. Chalupsky. Ontomorph: a translation system for symbolic knowledge. In A.G. Cohn, F. Giunchiglia, and B. Selman, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000), San Francisco, CA.* Morgan Kaufman, 2000.

[3] H. Chalupsky, E. Hovy, and T. Russ. Progress on an automatic ontology alignment methodology, 1997. ksl-web.stanford.edu/onto-std/hovy/index.htm.

[4] Michael Erdmann and Rudi Studer. Ontologies as conceptual models for xml documents. In *Proceedings of the 12th workshop on Knowledge Acquisition, Modeling, and Management*, 1999. At sern.ucalgary.ca/KSI/KAW/KAW99/papers.html.

[5] J. Euzenat. An infrastructure for formally ensuring interoperability in a heterogenous semantic web. In *Proceedings of the First Semantic Web Working Symposium, Palo Alto*, pages 345–360, 2001.

[6] FGDC. Spatial data transfer standard. Federal Geographic Data Committee, U. S. Geological Survey. Proposed standard, 1998.

[7] W. E. Grosso, H. Eriksson, R. W. Fergerson, J. H. Gennari, S. W. Tu, and M. A. Musen. Knowledge modeling at the millennium (the design and evolution of Protege-2000). Technical report, Stanford University, Institute

for Medical informatics, Stanford, CA, 1999. Technical Report SMI-1999-0801.

[8] Ian Horrocks, Frank van Harmelen, Peter Patel-Schneider, Tim Berners-Lee, Dan Brickley, Dan Connolly, Mike Dean, Stefan Decker, Dieter Fensel, Pat Hayes, Jeff Heflin, Jim Hendler, Ora Lassila, Deb McGuinness, and Lynn Andrea Stein. DAML+OIL, March 2001. www.daml.org/2001/03/daml+oil-index.html.

[9] E.H. Hovy. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC). Granada, Spain*, 1998.

[10] Jane Hunter. Adding multimedia to the Semantic Web - building an MPEG-7 ontology. In *Proceedings of the First Semantic Web Working Symposium, Palo Alto*, pages 261–284, 2001.

[11] International Hydrographic Organization. IHO transfer standards for digital hydrographic data, edition 3.0, 1996.

[12] C. A. Knoblock and J. L. Ambite. Agents for information gathering. In J. M. Bradshaw, editor, *Software Agents*, chapter 16. MIT Press, Cambridge, MA, 1997.

[13] Elbert S. Maloney. *Chapman Piloting: Seamanship and Boat Handling*. Hearst Marine Books, New York, 63rd edition, 1999.

[14] R. M. Malyankar. Acquisition of ontological knowledge from canonical documents. In *IJCAI-2001 Workshop on Ontology Learning*. Seattle, WA, 2001.

[15] D. McGuiness, R. Fikes, J. Rice, and S. Wilder. An environment for merging and testing large ontologies. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000), Breckenridge, Colorado*, April 2000. Tech. report KSL-00-16, Knowledge Systems Laboratory, Stanford University.

[16] Prasenjit Mitra, Gio Wiederhold, and Stefan Decker. A scalable framework for the interoperation of information sources. In *Proceedings of the First Semantic Web Working Symposium, Palo Alto*, pages 317–329, 2001.

[17] National Oceanic and Atmospheric Administration. Chart no. 1: Nautical chart symbols, abbreviations, and terms, 1997.

[18] N. F. Noy and M. Musen. SMART: Automated support for ontology merging and alignment. In *Twelth Workshop on Knowledge Acquisition, Modeling, and Management, Banff, Canada*, 1999.

[19] N. F. Noy and M. A. Musen. PROMPT: Algorithm and tool for automated ontology merging and alignment. Technical report, Stanford University, Institute for Medical informatics, Stanford, CA, 2000. Technical Report SMI-2000-0831.

[20] J. Spalding and A. Pirzada. Providing marine navigation information in the new millennium, 2001. IAIN World Congress, San Diego.

[21] J. Spalding, K. Shea, M.J. Lewandowski, and M. Fitz-Patrick. Intelligent waterway system and the waterway information network, 2002. Institute of Navigation National Technical Meeting, San Diego. (To appear.).

[22] James Hendler Tim Berners-Lee and Ora Lassila. The semantic web. *Scientific American*, May 2001.

[23] W3C (World Wide Web Consortium). Extensible markup language (XML) 1.0 (second edition), 2000. http://www.w3c.org/TR/2000/REC-xml-20001006.

[24] W3C (World Wide Web Consortium). Resource Description Framework (RDF) schema specification 1.0, 2000. At http://www.w3.org/TR/2000/CR-rdf-schema-20000327/.

[25] W3C (World Wide Web Consortium). XML schema part 0: Primer, 2001. http://www.w3c.org/TR/xmlschema-0/.

[26] W3C (World Wide Web Consortium). XML schema part 1: Structures, 2001. http://www.w3c.org/TR/xmlschema-1/.

[27] W3C (World Wide Web Consortium). XML schema part 2: Datatypes, 2001. http://www.w3c.org/TR/xmlschema-2/.